WPS4984

Policy Research Working Paper 4984

Diasporas

Michel Beine
Frédéric Docquier
Çağlar Özden

The World Bank
Development Research Group
Trade and Integration Team
July 2009

Policy Research Working Paper 4984

Abstract

Migration flows are shaped by a complex combination of self-selection and out-selection mechanisms. In this paper, the authors analyze how existing diasporas (the stock of people born in a country and living in another one) affect the size and human-capital structure of current migration flows. The analysis exploits a bilateral data set on international migration by educational attainment from 195 countries to 30 developed countries in 1990 and 2000. Based on simple micro-foundations and controlling for various determinants of migration, the analysis finds that diasporas increase migration flows, lower the average educational level and lead to higher concentration of low-skill migrants. Interestingly, diasporas explain the majority of the variability of migration flows and selection. This suggests that, without changing the generosity of family reunion programs, education-based selection rules are likely to have a moderate impact. The results are highly robust to the econometric techniques, accounting for the large proportion of zeros and endogeneity problems.

This paper—a product of the Trade Team, Development Research Group—is part of a larger effort in the department to understand the impact of international migration on poverty and development. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at cozden@worldbank.org.

Diasporas*

Michel Beine[a], Frédéric Docquier[b] and Çağlar Özden[c]

[a] University of Luxembourg and CES-Ifo

[b] FNRS and IRES, Université Catholique de Louvain, IZA-Bonn and CReAM-London.

[c] World Bank, Development Research Group

*Earlier versions of this paper have been presented at the "Migration and Development" conference (Lille, June 2008), at the "Globablization and Brain Drain" conference (Tel Avid and Jerusalem, December 2008). The paper benefitted from comments and suggestions by Luisito Bertinelli, Serge Coulombe, Caroline Freund, Eric Gould, Gordon Hanson, Will Martin, David McKenzie, Mario Piacentini, Samaschwar Rao, Hillel Rapoport, Assaf Razin, Mark Rosenzweig, Maurice Schiff and Antonio Spilimbergo. We would like to thank Sara Salomone for gathering data on guest workers' agreements. The second author acknowledges financial support from the Belgian Federal Government (PAI grant P6/07 Economic Policy and Finance in the Global Equilibrium Analysis and Social Evaluation) and the TOM (Transnationality of Migrants) Marie-Curie research and training network. The findings, conclusions and views expressed are entirely those of the authors and should not be attributed to the World Bank, its executive directors or the countries they represent.

"On the day I left Nigeria, I felt sad because I was leaving my family behind. I believed I would return eight years later, probably marry an Igbo girl, and then spend the rest of my life in Nigeria But 25 years ago, I fell in love with an American girl, married her three years later, and became eligible to sponsor a Green Card visa for my 35 closest relatives, including my parents and all my siblings, nieces and nephews. The story of how I brought 35 people to the United States exemplifies how 10 million skilled people have emigrated out of Africa during the past 30 years. We came to the United States on student visas and then changed our status to become permanent residents and then naturalized citizens Our new citizenship status helped us sponsor relatives, and also inspired our friends to immigrate here." (Philip Emeagwali)[1]

1 Introduction

Diasporas constitute invisible nations that reside outside their origin countries. In 2000, there were over 6 million Mexicans working in the United States, more than 1.2 million Turks in Germany and more than 0.5 million Algerians in France. In relative terms, 45 percent of the Surinamese-born were in the Netherlands; about 35 percent of the native-born from Grenada were in the United States; over 25 percent of Samoans were in New Zealand. Despite some of these staggering numbers, migrant diasporas exhibit diverse patterns, especially in terms of their human capital and education levels. Only 6.5 percent of the 22,000 Angolans in Portugal have post-secondary education whereas this proportion rises to 80 percent among the 715 Angolans in Canada. In total, 90 percent of all Angolan migrants with post-secondary education live in just five destination countries in the OECD.

This paper explores the role of existing diasporas on the size, educational structure and concentration of migration flows across different destinations. Understanding the role of migrant diasporas, especially how that role interacts with governments' migration policies is a critical issue for both sending and receiving countries. In addition to the welfare of its citizens living under other countries' jurisdiction, sending countries' governments are concerned about the costs and benefits of migration on the residents who stay at home. For the receiving countries, migrants generate significant externalities on the natives through capital and labor markets and as well public

[1]Extract of the keynote speech by Philip Emeagwali at the Pan African Conference on Brain Drain, Elsah, Illinois on October 24, 2003. Philip Emeagwali won the 1989 Gordon Bell Prize, which has been called "supercomputing's Nobel Prize", for inventing a formula that allows computers to perform their fastest computations - a discovery that inspired the reinvention of supercomputers. He was extolled by then U.S. President Bill Clinton as "one of the great minds of the Information Age" and described by CNN as "a Father of the Internet". He is the most searched-for scientist on the Internet.

finance channels (see Borjas, 1994, 1995, 1999, Razin and Sadka, 2004, Friedberg and Hunt, 1995, among others). In short, regardless of the question at hand, diasporas influence the welfare of all parties concerned - families back at home in the origin country, potential migrants searching for better opportunities and the natives in the destination country.

A large literature in sociology and economics has identified that migrants' networks facilitate further migration of people, movement of goods, capital, and ideas across national borders (see Rauch and Casella, 1998, Rauch and Trindade, 2002, Munshi, 2003, Rauch, 2003, Gao, 2003, Rapoport and Kugler, 2006, Docquier and Lodigiani, 2008). As it is presented repeatedly in the literature, the structure and the size of migration flows arise from a complex mix of self-selection factors (wage differentials, probability to find a job, welfare programs and amenities, migration costs, etc.) and out-selection factors (immigration policies at destination, mobility agreements, etc.). Our contribution is to show the role played by existing diasporas in shaping various characteristics of these flows.

Several studies focused on the self-selection mechanism, generally disregarding network externalities. Extending Roy's model (see Roy, 1951), Borjas (1987) demonstrate that migrants from poor countries with high returns to skills tend to be negatively selected, thus explaining how changes in the origin mix of US immigrants (from EU countries to Latin American and Asian countries) over time has affected their average skills and performance in the US labor market. Assuming that migration costs decrease with educational attainment, Chiquiar and Hanson (2005) develop a model compatible with positive, negative and intermediate selections, depending on the range of the schooling distribution. They find that Mexican emigrants, while much less educated than U.S. natives, are on average more educated than residents of Mexico and tend to occupy the middle and upper portions of Mexico's wage distribution. In terms of observable skills, there is intermediate or positive selection of immigrants from Mexico.

Existing migrant networks play an important role on the migration decisions of potential migrants. Relying on the informational and financial support provided by the network, newcomers can lower their migration and assimilation costs. As discussed in Massey et al. (1993), models of migrant diasporas are based on the theory of 'network externalities'. In particular, Carrington, Detragiache and Vishwanath (1996) show that when moving costs decrease with the size of the network already settled in the destination (an assumption which is supported by many sociological studies), migration occurs gradually over time. Migration tends to follow geographical, cultural or political channels and low-moving-cost individuals migrate first. Their presence lowers the migration costs of the next group and the process continues as long as benefits exceed costs of migration[2]. In addition to these cost-based network externalities, diasporas attract new migrants via family reunification programs if the

[2]Pedersen, Pytlikova and Smith (2008) also find evidence of strong network effects in immigration flows into 27 OECD countries during the period 1990-2000

destination country government has implemented them. In most continental European countries, family reunification is the main route for many potential migrants. Even in one of the most selective country such as Canada, about 40 percent of immigrants come under the family reunification and refugee programs, rather than selective employment or skill-based programs. Emegwali's quotation perfectly illustrates these channels. Through network effects ("our presence [...] inspired our friends to immigrate here") and family reunification programs ("I became eligible to sponsor 35 relatives for a Green Card"), existing diasporas positively impact future flows of migrants.

Only a few papers analyze the linkages between diasporas and the structure of migration flows. Building on Chiquiar and Hanson (2005), Mc Kenzie and Rapoport (2007) start from the intermediate selection case (which reflects the Mexico-to-US pattern) and demonstrate that a decrease in migration costs generally has a stronger effect on low-skill migration than on high-skilled migration.[3] Using survey data from Mexico, they show that the probability of migration increases with education in communities with low migrant networks, but decreasing with education in communities with high migrant networks. Taking advantage of a recent data set on international migration by educational attainment (see Docquier, Lowell and Marfouk, 2009), our paper generalizes this result by analyzing the role of diaspora size on the educational structure of migration from 195 countries to the 30 OECD countries. Accounting for the usual determinants of migration and correcting for several econometric problems, we show that larger diasporas increase migration flows and lower their average educational level, as expected. To reinforce this result, we analyze the effect of diasporas on the geographic concentration of high-skill and low-skill migrants. We show that diasporas increase the concentration of low-skill migrants relative to high-skilled ones.

Interestingly, diasporas explain a large portion of the variability of migrants' flows (71 percent) and selection (47 percent). These percentages capture both network externalities that lower migration costs and the effect of family reunification programs. Thus, without changing the generosity of these family reunion programs, education-based migrant selection rules are likely to have a moderate impact,especially in countries hosting large diasporas. These results are highly robust to various econometric techniques, accounting for the large proportion of zeros and possible correlation of the network size with unobservable components of the migration flows.

The remainder of the paper is organized as following. Section 2 describes migration data and presents some stylized facts on the size and structure of diaspora and migration flows. Section 3 derives testable predictions from a stylized theoretical model. Econometric issues and empirical results are presented in Sections 4 and 5. Finally, Section 6 concludes.

[3]Bertolini (2009) provides also similar evidence from the Ecuadorian migration to Spain and the US. The negative selection of Ecuadorian migrants to the US is largely explained by the size of the networks at destination.

2 Stylized facts

The term diaspora (in ancient Greek, "a scattering or sowing of seeds") refers to dispersion of any people or ethnic population, voluntarily or by force, from their traditional homelands and the ensuing developments in their culture in the destination, mostly as a minority. In the economic sense, the diaspora refers to migrants who gather in relatively significant numbers in a particular destination country or region. Some examples are the Turkish Gastarbeiter in Germany, South Asian workers in the Persian Gulf and Cuban migrants in the US.

Following this definition, we consider the size of a diaspora as the population (aged 25+) born in country i and living in country j. We use the Docquier, Lowell and Marfouk (2007, referred to as DLM from now on) database which extends and updates Docquier and Marfouk (2006). Based on census and register information on the structure of migrant communities in all OECD countries in 1990 and 2000, DLM database provides the stock of immigrants from any given country in each of the OECD countries by education level. The dataset covers only the adult population aged 25 and over, thus excludes children and students who emigrate temporarily to complete their education. In addition, migration is defined on the basis of the country of birth rather than citizenship[4].

The main strength of the DLM database is that it distinguishes between three levels of education for migrants. High-skilled migrants are those with post-secondary education. Medium-skilled migrants are those with upper-secondary education completed. Low-skilled migrants are those with less than upper-secondary education, including those with lower-secondary and primary education or those who did not go to school. The main characteristics of the diaspora that we consider in this paper are the following:

- The size of the diaspora, measured as the population aged 25+ born in country i and living in the OECD country j ($\neq i$).
- The education level of the diaspora, proxied by the log-ratio of the proportions of high-skill to low-skill migrants.
- The concentration of the diaspora, measured as the Herfindhal index applied to the distribution of the diaspora across different destinations.

Table 1 shows the 20 largest bilateral migrant communities residing in the OECD countries, both by overall size and by different education levels. The distinction

[4]Even though this is the standard definition of a migrant, especially in the economics literature, the dataset does not include second generation children who are born in the destination country even though they might constitute an important part of a diaspora in the sociological sense. This is simply due to absence of comprehensive administrative data in tracking of the migrants' children. However, we expect diaspora sizes inclusive and exclusive of second generation to be highly correlated.

between skilled and unskilled diasporas and its consequences is one of the most important contributions of this paper. With respect to the size, Table 1 allows to observe directly some of the determinants of the size of the diaspora, especially at a given destination country. As clearly seen in Table 1, the sizes of sending and receiving countries' populations are primary determinants of the size of the diasporas. That is why the United States appears as the home to many of the largest migrant communities and larger developing countries (such as Mexico, Turkey, the Philippines and India) are the main sending countries. Other factors, such as wage differentials, physical distance, linguistic proximity, colonial links, immigration policies at destination, are also frequently identified in the empirical literature as determinants of migration and clearly influence the migration corridors listed in Table 1.

In order to shed some preliminary light on how existing networks affect migration flows and and especially their human capital (educational) composition, let us look at the size and the educational structure of the Turkish diaspora in three different European countries: Germany, Spain and Luxembourg. Turkey is an interesting case since it does not have any colonial links, has no linguistic proximity with any of the major destination countries[5] but has large diasporas in a limited number of countries like Germany (see Table 1). The geodesic distance between Turkey and the three considered European countries is broadly the same and wage levels at destination are not very different across destination countries (they are higher in Luxembourg and lower in Spain). The data on the size of diaspora and the educational structure of those diasporas display striking differences. In 2000, there were only 194 Turkish migrants in Luxembourg, with 44% (26%) with a tertiary (primary) education level. In Germany, the corresponding figures are 1.2 million Turkish migrants with 6% (86%) with a tertiary (primary) educational level. In Spain, there were 1,040 Turkish migrants, with 33% (29%) with a tertiary (primary) educational level. This simple example highlights the striking relationship between migrants' networks and both the size and the skill composition of migration flows.

What is the extent of the relationship between diasporas and migration flows and how general is it in the data? Figure 1 provides another perspective and depicts the size of bilateral diasporas and the proportion of post-secondary educated (high-skilled) from four origin countries: Mexico, Morocco, Algeria, Mauritania. The curves are the exponential trends estimated for all origin countries and show that there is negative relationship between the diaspora size and the level of education. This figure shows the importance of analyzing bilateral data with econometric models that account for origin and destination country specific effects.

The next question is on the concentration/dispersion of migrants across different destinations. Figure 2 compares the concentration index (measured by the Herfindhal's index) of high-skill and low-skill migrants and indicates that there is a positive relationship between the two. In other words, for many source countries, both the

[5]Turkish is an Ural-Altaic language. The only European languages that are grammatically close are Finnish and Hungarian but they have almost no common vocabulary.

high and low skilled migrants tend to be either concentrated in few destination countries or relatively dispersed across the globe. A closer look also reveals that a larger share of the observations lie below the 45-degree line on the right side of the figure indicating low-skill migrants are even more concentrated than high-skilled migrants if the overall migration is concentrated. On the other hand, more observations on the left side of the figure are above the 45-degree line implying high-skill migrants are more concentrated if the overall concentration level is low. Another contribution of the paper is to empirically identify the determinants of the relative concentration (skilled vs unskilled) of the diasporas.

Table 1. Top-20 largest bilateral diasporas

Total diasporas			Highly skilled diasporas			Low skilled diasporas		
Origin	Destination	Size	Origin	Destination	Size	Origin	Destination	Size
Mexico	Un. States	6,374,825	Mexico	Un. States	919,139	Mexico	Un. States	4,454,823
Turkey	Germany	1,272,000	Philippines	Un. States	833,958	Turkey	Germany	1,097,000
Philippines	Un. States	1,163,555	India	Un. States	664,406	Portugal	France	493,459
Un. Kingdom	Australia	969,004	Canada	Un. States	439,163	Algeria	France	430,941
China	Un. States	841,699	Korea	Un. States	437,264	El Salvador	Un. States	393,157
India	Un. States	836,780	China	Un. States	434,547	Italy	Germany	367,000
Vietnam	Un. States	807,305	Un. Kingdom	Un. States	418,794	Morocco	France	336,375
Cuba	Un. States	803,500	Germany	Un. States	387,067	Cuba	Un. States	330,418
Canada	Un. States	715,825	Un. Kingdom	Australia	381,348	Italy	France	330,380
Korea	Un. States	676,640	Un. Kingdom	Canada	365,420	Vietnam	Un. States	310,608
Germany	Un. States	646,815	Vietnam	Un. States	347,127	China	Un. States	280,422
Un. Kingdom	Un. States	637,584	Cuba	Un. States	307,541	Dom. Rep.	Un. States	275,017
El Salvador	Un. States	619,185	Taiwan	Un. States	220,280	Spain	France	267,219
Un. Kingdom	Canada	580,250	Japan	Un. States	202,300	Guatemala	Un. States	218,124
Portugal	France	536,236	Jamaica	Un. States	199,321	Bulgaria	Turkey	211,172
Dom. Rep.	Un. States	527,520	Colombia	Un. States	184,472	Italy	Un. States	206,460
Algeria	France	512,778	Poland	Un. States	182,300	Italy	Canada	200,665
Italy	Un. States	461,085	Iran	Un. States	174,043	Un. Kingdom	Australia	191,764
Italy	Germany	456,000	Russia	Un. States	156,984	Ukraine	Poland	190,578
Jamaica	Un. States	449,795	Philippines	Canada	154,960	India	Un. Kingdom	178,551

Source: Docquier, Lowell and Marfouk (2009)

Figure 1. Percentage of highly skilled (Y-Axis) and Log size (X-axis) of diasporas for selected countries

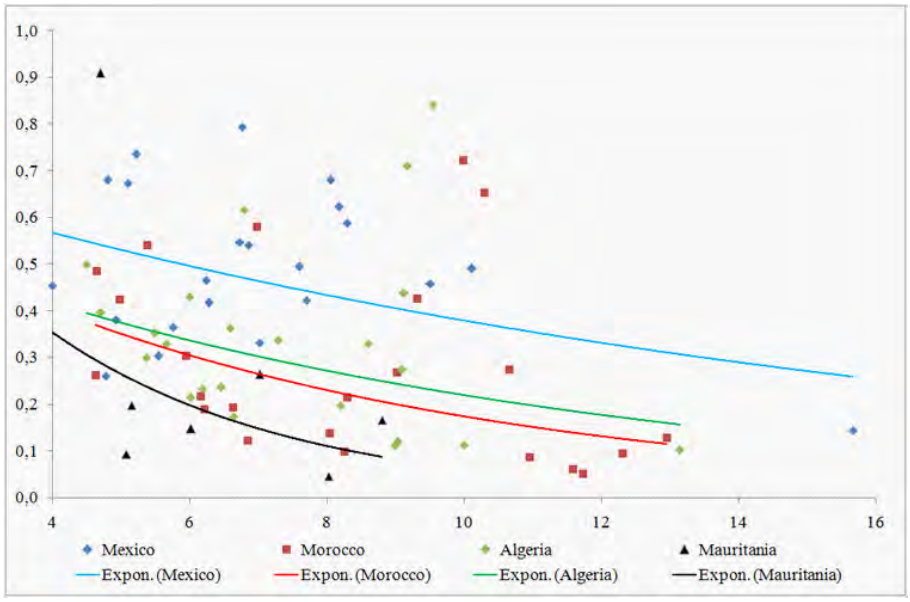
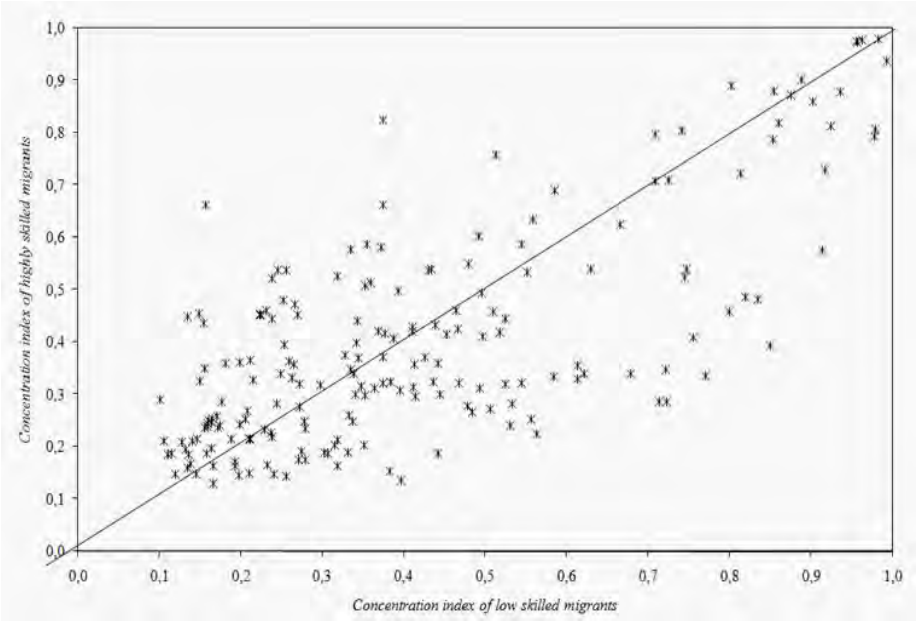


Figure 2. Concentration of the high-skilled (Y-axis) and low-skilled (X-axis) diasporas



3 Theoretical foundations

We consider model of migration with a single skill type in order to model the effects of diasporas. A worker endowed with h units of human capital earns a wage $w_i h$ in country i where w_i is the skill price in that country. As in Rosenzweig (2008), this structure reflects the assumptions that (i) the main source of variation in wages *within* a country is the differences in the human capital levels (h) of the residents and (ii) the source of variation in wages *across* countries is the differences in average skill levels and skill prices (w_i). The individual utility is linear in income but also depends on possible moving costs and characteristics of the country of residence. The utility of a type-h individual born in country i and staying in country i is given by:

$$u_{ii}(h) = w_i h + A_i + \varepsilon_i$$

where A_i denotes country i's characteristics (amenities, public expenditures, climate, etc.) and ε_i is a iid extreme-value distributed random term. The utility obtained when the same person migrates to country j is given by

$$u_{ij}(h) = w_j h + A_j - C_{ij}(.) - V_{ij}(.) + \varepsilon_j$$

The migration costs are divided into two categories. C_{ij} captures moving and assimilation costs that are borne by the migrant. These would include transportation costs, expenditures to learn the new language, find a job and obtain necessary licences to practice a profession etc. V_{ij} represents policy induced costs borne by the migrant to overcome the legal hurdles set by the destination country's government's policies. These costs include visa fees, the bureaucratic barriers for citizenship or even the amount paid to smugglers above the normal cost of transportation when legal entry is restricted. For simplification, we slightly abuse the terminology and refer to C_{ij} as migration costs and to V_{ij} as visa costs. They both depend on the existing diaspora networks and human capital level of the migrant as explained below. The main motivation to differentiate between these two types of costs is to identify the role of government's policy on migration flows and characteristics.

Let N_i denote the size of the native population that is within migration age in country i. When the random term follows an iid extreme-value distribution, we can apply the results in McFadden (1974) to write the probability that a type-h individual born in country i will move to country j as

$$\Pr\left[u_{ij}(h) = \max_k u_{ik}(h)\right] = \frac{N_{ij}}{N_i} = \frac{\exp\left[w_j h + A_j - C_{ij}(h) - V_{ij}(h)\right]}{\sum_k \exp\left[w_k h + A_k - C_{ik}(h) - V_k(h)\right]}$$

Similarly, the ratio of emigrants in country j to residents (N_{ij}/N_{ii}) is given by the following expression

$$\frac{N_{ij}}{N_{ii}} = \frac{\exp\left[w_j h + A_j - C_{ij}(.) - V_{ij}(.)\right]}{\exp\left[w_i h + A_i\right]}$$

or, in logs,

$$\ln\left[\frac{N_{ij}(h)}{N_{ii}(h)}\right] = (w_j - w_i)\,h + (A_j - A_i) - C_{ij}(.) - V_{ij}(.) \tag{1}$$

The ratio of immigrants to different destinations (N_{ij}/N_{ik}) or migrants to the same destination with different human capital levels may be expressed using similar expressions.

Migration costs, C_{ij}, depend on factors such as physical distance ($d_{i,j}$), destination and origin countries' social, cultural and linguistic characteristics (x_i, y_j) as well as human capital level (h) of the migrant and the size of the diaspora abroad ($M_{i,j}$). Thus, we write

$$C_{ij}(h) = c(d_{ij}, M_{ij}, x_i, y_j; h) \tag{2}$$

Distance has a negative effect on migration so $c_d^{'} > 0$. Because social networks lower information, assimilation and adaptation costs, diaspora has a positive effect on migration and lowering of costs so $c_M^{'} < 0$. The assumption $c_h^{'} < 0$ captures the facts that skilled migrants are better informed than the unskilled, have higher capacity to assimilate or have more adaptive skills and, thus, face lower migration costs. Finally, we assume that the advantages of being skilled are likely to be more important when the diaspora size is small and migrants can not rely on others. When the diaspora size is larger, the cost advantages of being skilled decline, i.e. $c_{hM}^{''} > 0$.[6]

The legal (or the visa) costs, V_{ij}, are determined by the destination country j's government's policies and depend on various factors. These policies can be specific to sending country i or depend on individual characteristics of the migrants. Many destination countries have specific programs for family reunification or for highly skilled individuals. Other countries sign bilateral free mobility agreements or grant automatic citizenship based on colonial links, common ethnicity or religion. The green card lottery program of the US, for example, has country-specific quotas.

Diasporas affect the visa costs mainly through family reunification programs. Let f_j denote the generosity of the family reunification program of country j which generally does not discriminate between different origin countries. The probability that a potential migrant from country i has a relative in country j is an increasing function of M_{ij}/N_i. Thus, the overall effect of reunification programs on visa costs depends on the expression $\frac{f_j M_{ij}}{N_i}$.

The migrant's human capital level also affects the visa costs if there are selective immigration programs such as the H1-B program in the US. We denote the generosity of economic migration programs as e_j and the overall effect of human capital on visa costs depends on $e_j h$. Finally, we formalize the presence of free mobility agreements (such as those between EU members) through a dummy variable b_{ij} which is equal

[6] Analyzing the Mexican migration to the US, Mc Kenzie and Rapoport (2007) provide evidence that the decrease in migration costs due to the network effect is stronger for low skilled migrants.

to one if an agreement exists. As a result, we define visa costs as

$$V_{ij}(h) = (1 - b_{ij})v\left(\frac{f_j M_{ij}}{N_i}, e_j h\right) \tag{3}$$

Policy variables, f_j and e_j, only matter for origin countries that do not have free mobility agreements with country j (when $b_{ij} = 0$). The partial derivatives of $v(.)$ with respect to both of the arguments are negative, $v_f^{'} < 0, v_e^{'} < 0$, $v_{ff}^{''}, v_{ee}^{''} \gtrless 0$ and $v_{ef}^{''}(.) > 0$: the probability that an individual relies on family reunion program decreases (resp. increases) when economic program becomes more (resp. less) generous or *vice versa*.

The net effect of human capital level on visa costs is given by

$$\frac{\partial V_{ij}}{\partial h} = (1 - b_{ij})e_j v_e^{'}\left(\frac{f_j M_{ij}}{N_i}, e_j h\right) < 0,$$

The effect of human capital on visa costs also depend on the size of the diaspora. When the diaspora size is bigger, the probability that a migrant relies on an economic migration program declines and the probability he relies on family reunion programs increases. Hence, we have

$$\frac{\partial\left(\partial V_{ij}/\partial h\right)}{\partial M_{ij}} = (1 - b_{ij})e_j \frac{f_j}{N_i} v_{ef}^{''}\left(\frac{f_j M_{ij}}{N_i}, e_j h\right) > 0$$

since $v_{ef}^{''}(.)$ is positive. With these definitions in place, we can write (1) as

$$\begin{aligned} \ln\left[\frac{N_{ij}(h)}{N_{ii}(h)}\right] &= (w_j - w_i)\, h + (A_j - A_i) - c(d_{ij}, M_{ij}, x_i, y_j; h) \\ &\quad -(1 - b_{ij})v\left(\frac{f_j M_{ij}}{N_i}, e_j h\right) \end{aligned} \tag{4}$$

3.1 Self-Selection

This simple model and the underlying assumptions allow us to analyze major characteristics of diasporas, especially how the existing diaspora influences the size of migrant flows, their composition in terms of human capital and concentration across different destinations. Before proceeding to these questions, we first analyze how changes in human capital level influence the migration decision of the individual and the overall migration level. From equation (4), we have

$$\frac{\partial \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial h} = (w_j - w_i) - c_h^{'} - (1 - b_{ij})e_j v_e^{'} \tag{5}$$

which is positive if $-c_h^{'} - (1 - b_{ij})e_j v_2^{'}\left(\frac{f_j M_{ij}}{N_i}, e_j h\right) > w_i - w_j$[7].

In the case of South-North migration, we have $w_j > w_i$ and, therefore, above condition always holds. Hence, level of of migration increases with human capital levels and positive selection is observed. Positive selection is even stronger when network effects on moving costs $\left[-c_h^{'}\right]$ are large and when the host country has a selective immigration policy (i.e. e_j is large). We should note that positive selection does not imply that there are more skilled emigrants than unskilled emigrants, but the higher-skilled have a higher propensity to migrate. If the proportion of the highly-skilled among natives is low (such as in Africa), there will still be more unskilled than skilled migrants in destination countries. However, the ratio of the skilled to the unskilled will be higher among migrants when compared to natives. For other types of migration (between rich and rich, between poor and poor, or from rich to poor countries), we might have $w_j - w_i < 0$. In that case, negative selection could emerge.

3.2 Diaspora Externalities

We now turn to diaspora effects on the size and structure of migration flows. First, from (4), a large diaspora in destination j unambiguously increases current migration flows to j from i:

$$\frac{\partial \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial M_{ij}} = -c_M^{'} - (1 - b_{ij})\frac{f_j}{N_i}v_f^{'} > 0 \tag{6}$$

The overall impact depends on the effect of networks on migration costs ($c_M^{'}$) and on the generosity of family reunion programs (f_j) together with the effect on visa costs ($v_f^{'}$). Second, we show that a larger diaspora in country j reduce the 'positive selection' of migrants to j from i:

$$\frac{\partial^2 \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial h \partial M_{ij}} = -c_{hM}^{''} - (1 - b_{ij})e_j\frac{f_j}{N_i}v_{ef}^{''} < 0 \tag{7}$$

3.3 Immigration Policies

What are the implications of these results for immigration policies? Obviously, a more generous immigration policy, both in terms of family reunification and economic immigration programs, at destination increases the size of immigration flows:

$$\frac{\partial \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial f_j} = -(1 - b_{ij}).\frac{M_{ij}}{N_i}.v_f^{'} > 0 \tag{8}$$

$$\frac{\partial \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial e_j} = -(1 - b_{ij}).h.v_e^{'} > 0 \tag{9}$$

[7]In practice, some reported zeros might not reflect the actual absence of migrants. Due to confidentiality and disclosure rules, some statistics offices report a zero when the diaspora size is lower than a threshold value. We are not able to distinguish these cases from "true" zeros.

Immigration policies also affect the selection of immigrants. Since $v_{ef}^{''}$ is positive, stronger emphasis on family reunion programs (higher f_j) reduces the quality (i.e. the positive selection) of immigrants:

$$\frac{\partial^2 \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial h \partial f_j} = -(1-b_{ij}).e_j.\frac{M_{ij}}{N_i}.v_{ef}^{''} < 0$$

The effect of stronger economic migration programs (higher e_j) on the selection of immigrant is somewhat ambiguous since the first term of the expression below is positive and the second term is negative. A close inspection, however, shows that the net effect is likely to be positive unless $v_{ee}^{''}$ is strongly negative.

$$\frac{\partial^2 \ln\left[N_{ij}(h)/N_{ii}(h)\right]}{\partial h \partial e_j} = -(1-b_{ij}).v_e^{'} - (1-b_{ij}).e_j.h.v_{ee}^{''} \gtrless 0$$

Our simple model provides many interesting insights and gives rise to many testable predictions. Due to data availability (especially, in the absence of detailed data on bilateral immigration policies), we focus on some important predictions of the empirical section. These can be summarized as follows:

- The effect of diasporas on the migration flows is unambiguously positive. This impact is composed of the reduction of migration costs and visa costs through a stronger family reunification effect. Both effects yield a total positive impact.
- The effect of diasporas on the selection of migrants and the skill ratio is negative. A larger diaspora lowers migration and visa costs for all skill levels but the intensity of reduction is stronger for low-skilled migrants.
- The impact of diasporas on the concentration level should be in line with the effect in terms of selection. In particular, if diasporas tends to benefit a negative selection process, it should increase the concentration of low-skill migrants compared to the concentration of high-skill migrants.

4 Empirical Analysis

In this section, we analyze the determinants of the important characteristics of international migration flows - their size, their educational composition and their relative concentration by education level across different destination countries. In particular, in line with the theoretical model, we assess the impact of existing diasporas as well as other factors that influence migration flows. We start with OLS regressions but also account for important econometric problems using other techniques. The first important issue is the high proportion of observations with either zero or undefined

values[8]. The second one is the correlation between the diaspora size and the error term, due to the presence of some unobservable bilateral components that affect both the size of the diaspora and migration flows. One important aspect of the whole analysis is the robustness of the main results to alternative estimation techniques.

4.1 Size

The first question we ask is on determinants of migration flows and the role of the diaspora size. In equation (4), the dependent variable is $\ln[N_{ij}(h)]$, i.e. the log of the migration flow between 1990 and 2000 from country i to country j of individuals with skill level h. We proxy it by taking the difference of the migration stocks observed in 1990 and 2000.

Among the main determinants of migration flows in equation (4) are the wage differential (specific to each skill level), migration costs and the factors influencing visa costs and other legal barriers. In Appendix B, we report the data sources and the way we construct measure the explanatory variables that proxy determinants of migration flows. We have good estimates for skill prices in destination countries (w_j) but fairly imprecise data on wages at origin (w_i) in order to construct the wage differential variable ($w_j - w_i$). One way of resolving this problem is to include origin country dummies γ_i that capture the combined effect of all unobserved characteristics of the origin country i on the migration flow to country j. These origin country dummies also capture the role of stock of residents with education level h ($\ln[N_{ii}(h)]$) as well all migration costs specific to the origin country (x_i) in equation (4). Pair-specific factors influencing migration costs are captured by geographical distance between the two countries, colonial links (a dummy variable) and linguistic proximity. We also introduce a dummy variable indicating whether the two countries are subject to the Schengen agreement favouring the mobility of persons within the European Community. The set A_j includes destination-specific variables that affect the attractiveness of country j in terms of migration such as population sizea and social expenditures as a share of GDP (as a measure of the extent of social welfare). The proxy for selective immigration policies is measured by the share of refugees in immigrants admitted in 1990 by country j. Finally, we capture diaspora effects by size of the diaspora in 1990 and denoted by the variable $M_{i,j}$. It should be clear that the estimated impact of $M_{i,j}$ in the estimation is a combined effect through C_{ij} (network effects that lower migration costs) and the impact on V_{ij} (family reunification effects that lower visa costs).

Introducing these variables, we get a first specification for the migration flow with observable destination specific variables:

[8]Some reported zeros might not reflect the actual absence of migrants. Due to confidentiality and disclosure rules, some national statistics offices report zero when the diaspora size is below a threshold level. We are not able to distinguish these cases from "true" zeros.

$$\ln\left[N_{ij}(h)\right] = \alpha_0 + \alpha_1 \ln\left(M_{ij}\right) + \alpha_2 d_{i,j} + \alpha_3 w_j + \alpha_4 A_j + \gamma_i + \epsilon_{ij} \tag{10}$$

where $\ln\left[N_{ij}(h)\right]$is the change in the migrant stock observed between 1990 and 2000 from country i to country j with education level h, M_{ij} is the size of the diaspora in 1990, $d_{i,j}$ is a vector of other observable bilateral variables affecting the migration costs as described above, w_j is the level of wages at destination and A_j is a set of other destination specific variables thought to affect the attractiveness of country j.

Above specification assumes that the effect of all destination country specific variables is well captured by w_j and A_j. This is obviously a strong assumption as it is very likely that other factors play a significant role in attracting migrants in country j. In addiition, some variable such as the immigration policy might be measured in an imprecise way. The empirical measurement of immigration policies is a well known challenge in the literature and has so far not received a full satisfying treatment. Since we are mainly interested in estimating the impact of M_{ij} , in the next specification, we introduce destination country dummies γ_j that capture the combined impact of unobserved characteristics of host countries:

$$\ln\left[N_{ij}(h)\right] = \alpha_0 + \alpha_1 \ln(M_{ij}) + \alpha_2 d_{i,j} + \gamma_j + \gamma_i + \epsilon_{ij}. \tag{11}$$

Compared to the previous model in (10), introduction of destination country dummies lead to an improvement of the specification and thus can minimize the case of a misspecification bias. Our results in the next section show that insertion of destination fixed effects leads to an increase in the R^2 by more than 10 percents. This model should thus be preferred, at least as far the estimation of α_1 is concerned.[9]

4.2 Selection

We use the selection ratio, the number of skilled over unskilled migrants, as the proxy for educational (or the human capital) structure of migration flows and diasporas. It is defined as $S_{ij} = \frac{M_{ij}(s)}{M_{ij}(u)}$, where $M_{ij}(s)$ and $M_{ij}(u)$ refer to the number of skilled and unskilled migrants respectively. In line with Grogger and Hanson (2008) and the original definition in Docquier, Lowell and Marfouk (2007), we define skilled and unskilled migrants as migrants with post-secondary and primary education levels, respectively . Equation (4) can be manipulated to be written in terms of the ratio of different skill levels to the same destination as a result of the extreme-value assumption of the error term. Depending on the introduction of destination dummies or not, the estimated equations are :

$$ln(S_{ij}) = \alpha_0 + \alpha_1 \ln(M_{ij}) + \alpha_2 d_{i,j} + \alpha_3 w_j + \alpha_4 A_j + \gamma_i + \epsilon_{ij} \tag{12}$$

[9]Of course, the cost of adopting specification (11) is that, we can not estimate the impact of destination specific variables such as the wage levels w_j in host countries. Please refer to Rosenzweig (2008) and Grogger and Hanson (2008) for a discussion.

and

$$ln(S_{ij}) = \alpha_0 + \alpha_1 \ln(M_{ij}) + \alpha_2 d_{i,j} + \gamma_j + \gamma_i + \epsilon_{ij} \quad (13)$$

The availability of data for 1990 also allows us to study the impact of diaspora on the change in the selection ratio (which is broadly equal to the selection ratio of new migrants). The two estimated specifications are then obtained by substituting $ln(S_{ij})$ by its change between 1990 and 2000, $\Delta ln(S_{ij})$.

4.3 Relative Concentration

We also explore the relative concentration of diasporas across education levels. In particular, we ask whether diasporas tend to lead to more concentration of unskilled rather than skilled migrants at a given destination. We construct our destination-specific relative concentration measure as the following:

$$C^s_{ij} - C^u_{ij} = \left[M_{ij}(s) / \sum_i M_{ij}(s) \right]^2 - \left[M_{ij}(u) / \sum_i M_{ij}(u) \right]^2$$

where indices s and u refer to skilled and unskilled migrants. A nice property of this bilateral measure is that its sum across destination countries j boils down to the difference between Herfindhal indices for skilled and unskilled migrants.

Once again, we consider regression models with and without destination dummies and consider regression on levels (relative concentration $C^s_{ij} - C^u_{ij}$ observed in 2000) and on change between 1990 and 2000. The models for the levels are:

$$C^s_{ij} - C^u_{ij} = \alpha_0 + \alpha_1 \ln(M_{ij}) + \alpha_2 d_{i,j} + \alpha_3 w_j + \alpha_4 A_j + \gamma_i + \epsilon_{ij} \quad (14)$$

and

$$C^s_{ij} - C^u_{ij} = \alpha_0 + \alpha_1 \ln(M_{ij}) + \alpha_2 d_{i,j} + \gamma_j + \gamma_i + \epsilon_{ij} \quad (15)$$

The specifications relative to the changes are obtained by substituting $C^s_{ij} - C^u_{ij}$ relative to 2000 by $\Delta(C^s_{ij} - C^u_{ij})$ where Δ refers to the change between 1990 and 2000. The latter specification is particularly demanding since the dependent measures "a difference in differences" of concentration rates.

4.4 Econometric Issues

The estimation of models (10-15) entails several econometric challenges that might lead the estimation of those models by OLS to generate inconsistent estimates. There are two basic reasons. The first one is related to the occurrence of zero or undefined values for the dependent variables in a large portion of the observations. The second one is the potential correlation of $\ln(M_{ij})$ with ϵ_{ij} due to the presence of an unobservable component affecting the size of the diasporas and the characteristics of new migrants. We now discuss how we address these issues.

4.4.1 Zero or undefined values for dependent variables

One of the most important features of our dataset is the high proportion of zero observations either for the size of diasporas in 2000 or for the flows of migrants between 1990 and 2000. This naturally occurs in many migration datasets as there is almost none or minimal migration for many country pairs. Pooling the data across the two periods, we have zero values in about 31% of the observations for the stock of migrants and in around 36% for the flows.

Our model is fully consistent with such large number of zero observations. Predicting a continuous number of emigrants, our model is an approximation of the "discrete-number" real world with $N_{ij}(h) \in \mathbb{N}$. If $\ln[N_{ij}(h)] < 0$, less than one migrant wants to leave her country[10]. This means that the bilateral migration flow is nil. The probability that $N_{i,j}(h) = 0$ is

$$\Pr\left[(w_j - w_i)\, h + (A_j - A_i) - C_{ij} - V_{ij} + \ln\left[N_{ii}(h)\right] < 0\right]$$

This case might arise for a number of reasons such as low wage differentials, large distances, high migration or visa costs. In turn, those latter costs obviously depend on the size of the existing diaspora.

Large number of zero observations occurs frequently in other empirical studies in international economics such as gravity equations in trade models. In the estimation of models (10-11) by OLS for the size of migration flows, the high occurance of zero values is likely to lead to inconsistent estimates. The use of a log specification drops the zero observations from the sample which is likely to result in biased estimates of the impact of diasporas and other variables on the migration flows and their selection. For instance, it might be the case that there are no migrants from country i to country j because migration costs are too high. In turn, migration costs might be too high because distance is too high and there is no diaspora. In this case, the exclusion of those observations leads to underestimation of the impact of the variables affecting the migration costs such as distance, colonial links, linguistic similarities or diasporas.

The first alternative is to use Poisson regression models that relies on pseudo maximum likelihood estimates, as advocated by Santos Silva and Tenreyro (2006) who show that the use of log linearization for gravity models leads to inconsistent estimates of the coefficients (such as the one relative to distance). A first reason, as mentioned before, is the exclusion of zero observations for the dependent variable. A second reason is that the expected value of the error will depend on the covariates of the model and hence will lead to estimation biases of the coefficient. In order to address that, we carry out Poisson regressions of the models explaining the size of the migration flows (i.e. models 10-11). The Poisson solution is nevertheless unfeasible for the selection and the concentration analyses. For the selection, the existence of

[10] In practice, some reported zeros might not reflect the actual absence of migrants. Due to confidentiality and disclosure rules, some statistics offices report a zero when the diaspora size is lower than a threshold value. We are not able to distinguish these cases from "true" zeros.

zero values for $M_{i,j}(h)$ leads to undefined values for S_{ij}, which cannot be handled by the Poisson approach.[11] For the concentration regressions, we end up with many negative values (more concentration for the unskilled compared to the skilled), which precludes the use of Poisson regression since they are count data models.[12]

A second alternative involves techniques accounting explicitly for a potential selection bias by two-step Heckman regression. In general, for all the features that we analyze (migration flows, selection and relative concentration), the first step involves the estimation of a selection equation - the probability for a given country pair to have a positive migration flow.[13] The usual procedure implies the use of an instrument in the probit equation, i.e. a bilateral variable that influences the probability of observing a diaspora between the two countries but does not influence the size of this diaspora.

It is difficult to find such an instrument but one possible candidate is diplomatic representation of the destination country in the origin country. Diplomatic representation might affect the probability of having at least one migrant by setting some kind of threshold on the initial migration and visa costs faced by potential migrants. In the absence of any diplomatic representation of country j in country i, the cost to get a visa can simply be too high so that nobody would consider to migrate to country j. The role of diplomatic representation in the migration process is to a certain extent analogous to the role played by a common religion for trade relationships. As argued by Helpman et al.(2007), a common religion (a proxy of costs of establishing business linkages) affects the extensive margin of trade (i.e. the probability of export) but not the intensive margin (i.e. trade volumes). In regressions (10-13), the use of a two-step Heckman approach yields intuitive results both for the flow and for the selection equation. In particular, for the selection equation, we find that diplomatic

[11] Strictly speaking, the estimation of models (12-13) leaves out a set of observations for two reasons. The main reason is that the selection ratio is undefined due to the fact that $M_{ij}(u) = 0$, i.e. the size of the unskilled diaspora is equal to zero. Poolling the data across the time periods, the fact that there is no unskilled diaspora leads to the exclusion of 35.7% of the observation. A second minor reason is that the use of the log of the skill ratio leaves out observations for which we observed $M_{ij}(s) = 0$ and $M_{ij}(u) > 0$, i.e. a diaspora with some unskilled migrants but no skilled migrants. The log transformation leads to a further exclusion of 256 pairs of countries (for 1990 and 2000), i.e. to an additional exclusion of 2.1% of the total observations.

[12] For the relative concentration, we could include in the OLS regressions zero values. Nevertheless, in order to have consistent subsamples with the analysis of selection and size, we consider a subsample of pairs for which we have non zero values for $C^s_{ij} - C^u_{ij}$. These zero values are exclusivelyrelated to zero values for both concentration indexes , i.e. correspond to $C^s_{ij} = 0$ and $C^u_{ij} = 0$. In other words, we have no case for which concentration levels would be positive and exactly similar between skilled and unskilled.

[13] To be more precise, for the analysis of migration stock, the probability that a given observation will be included in the regression is directly related to the probability of observing a diaspora (either regardless of the skill level, either for a particular skill level) for this country pair. For the migration flows, the probability is exactly the same since we have no case of zero migration flow with positive values of the stock in 1990 and 2000. For the analysis of selection, the probability is related to the existence of a diaspora or at least a skilled diaspora.

representation of country j in county i tends to positively affect the probability of observing a diaspora of country i in country j. Furthermore, the mills ratio turns out to be significant in the flow equation, suggesting that accounting for a selection bias is important.

Since the observed level of diaspora in 1990 is used as a regressor, the use of diplomatic representation leads to some collinarity problems in the selection equation. In order to mitigate the collinearity problems, it is possible to run Heckman two-step regressions without any additional instrument. As stressed by Wooldridge (2002), the use of an additional instrument in the probit equation is not strictly necessary. The drawback of not using an additional instrument is that the Mills ratio might become highly collinear with the explanatory variables of the flow equation, which in turn lowers the significance of the coefficients. This is not the case for most of our regressions. This method will therefore be used in the benchmark regressions. Nevertheless, as a robustness check, we carry out the same regressions using diplomatic representation as an instrument (Appendix A).

4.4.2 Correlated unobservables with the diaspora

One issue in identifying and estimating endogenous social effects (like the network effects in this paper) is the presence of unobservable correlated effects as explained by Manski (1993). In our framework, it could be the case that unobservable bilateral components affect the size of the diaspora M_{ij} and the dependent variables. For instance, unobserved cultural proximity between country i and country j might affect simultaneously the stock of migrants, the current flows of new migrants and their selection. The cross-sectional nature of the data prevents us to estimate directly those unobservable components. Therefore, those effects will be included in the error term, which in turn leads to some kind of omitted variable bias and to some correlation between M_{ij} and the error term.

We follow Munshi (2003) and proceed to a variable instrumental estimation of model (11) and (13) in order to address this issue and check the robustness of the results. In each case, we consider two instruments, i.e. variables correlated with M_{ij} but uncorrelated with the migration flows or the selection ratio. The use of two instruments allows us to check the empirical validity of this second condition through Hansen over-identification tests. Our first instrument is a dummy variable capturing whether the two countries were subject to a temporary guest worker agreement in the 1960s and 1970s. One can expect those guest worker agreements to exert a strong impact on the initial formation of a stock of migrants in the 1960s and the 1970s, hence influencing the stock in 1990. In contrast, it is unclear why those initial agreements would influence the contemporaneous migration flows beyond the impact exerted by the diaspora itself. Examples of such a process are illustrated by the impact of the post-war guest worker agreements between Belgium and Italy or Spain.

The second instrument is a variable capturing the unobserved diaspora in the

1960s through a combination of variables representing some push factor in country i, size in country i, openness and size in country j and distance between i and j. The basic measure is

$$IV_{ij} \equiv \ln(pop_i * immst_j / dist_{ij}) * confl_i$$

where pop_i is the population size in the 1960s of country i, $immst_j$ is the immigration stock of country j in the 1960s, $dist_{ij}$ is the distance between i and j and $confl_i$ is a dummy variable capturing the occurrence of armed conflicts in country i during the 1960s.

Our instrument should be correlated with the size of the diaspora observed in 1990. The variable pop_i is used as a proxy for the size of potential migrants in sending country i while $immst_j$ is a proxy of the openness and the size of the receiving country j in the 1960s. The product of the two is divided by the distance between the two countries captures the size of migration costs. This variable is multiplied by the conflict variable specific to the sending country to capture push factors causing people to leave country i. If this last variable is not correlated too much over time, this should impact the stock of migrants in the 1960s but not the flows of migrants coming from country i in subsequent periods such as the 1990-2000 period. In other terms, the low degree of serial correlation in the $confl$ variable ensures that our IV_{ij} variable is uncorrelated with our dependent variable, as the usual over identification test supports the exclusion restriction.

We only consider conflicts observed between 1946 and 1960 in order to capture push-factors leading to emigration in the 1950s and 1960s. We distinguish minor conflicts (number of battle-related deaths between 25 and 999) denoted CONFL1 and wars (at least 1,000 battle-related deaths in a given year) denoted CONFL2. We first use CONFL1; then we use CONFL2 and finally we add up the two variables. F-stat statistics of first stage regressions show that the correlation between this instrument set and the diaspora is relatively high. The results of the Hansen over-identification test suggest furthermore that the second condition of no correlation between the instrument set and the error term is supported by the data.

5 Estimation Results

5.1 Impact on Flows

Table 2 presents the estimation results regarding the determinants of migration flows and especially the role of diasporas. Columns (1) through (4) report the results on aggregate flows while columns (5) through (8) give the results for low-skilled and high-skilled migration flows. The OLS estimates of equations (10) and (11) are presented in columns (1) and (2) where a significant number of observations with zero migration flows (and the size of the diaspora in 1990) are dropped. Columns (3) through (8) report the results from the two-step Heckman approach where the regressions without additional instruments are used as the benchmark.

In appendix A, we check the robustness of the results presented in Table 2. Table A1 presents the coefficients obtained with using diplomatic representation as an additional instrument in the Heckman regressions. Table A2 presents the results of the Poisson regressions. As expected, we find that diplomatic representation significantly increases the probability of having a diaspora at destination, reflected by the positive and significant coefficient in the selection equation. In all Heckman regressions, the Mills ratio is statistically significant, which suggests that dealing with the large number of zero observations is important. However a comparison of the results in Table 2 with those in Tables A1 and A2 reveals that the estimated coefficient of the lagged diaspora variable is strikingly robust across estimation methods. Therefore, we focus on Table 2 in discussing the implications of our results.

Migration costs, as captured by bilateral distance and linguistic proximity variables, are found to exert significant effects on the migration flows whereas Schengen agreement seems to favor migration of highly skilled workers. Besides those predictable results, the effect of diasporas on the migration flows is quite important with a positive and significant coefficient. In the case with both destination and origin dummies, this coefficient lies between 0.62 and 0.77. Note that the specification used in (10-11) is similar to that of a β-convergence model. A positive coefficient for the lagged diaspora implies that there is no sign of convergence in the size of bilateral stocks of migrants, even when controlling for country fixed effects (capturing populations, individual domestic policies and economic conditions that influence incentives to migrate). This is probably due to the fact that migration to the North, especially from the South, sharply increased during the nineties. Since our period of interest is 1990-2000, our results clearly illustrate that country pairs with large initial diasporas exhibit higher growth rates compared to pairs with smaller diasporas.

As expected, OLS leads to an underestimated coefficient due to the exclusion of zero observations and the related selection bias. Methods that account for those zero values lead to slightly higher estimates. The estimated coefficient is almost the same in the Heckman two-step and Poisson regressions, emphasizing the robustness of the results. It is also quite similar whether we include an instrument in the selection equation of the two-step Heckman approach (Table A1)[14].

[14]With diplomatic representation used as an instrument in the selection equation, we get a coefficient of 0.660 for the impact of diaspora instead of 0.699 in the benchmark regressions. Note that the difference is not exclusively due to the estimation method since the use of diplomatic representation causes a loss of additional observations (190 origin countries instead of 195 in the benchmark regressions).

Table 2. Determinants of migration flows by skill level : OLS and Heckman regressions (1/2).

	(1) Total	(2) Total	(3) Total	(4) Total	(5) Low-skill	(6) Low-skill	(7) High-skill	(8) High-skill
Lagged diasp	0.620	0.616	0.699	0.831	0.778	1.192	0.625	0.728
	(34.35)***	(26.60)***	(43.91)***	(23.44)***	(22.25)***	(6.90)***	(44.57)***	(28.29)***
Col links	0.331	0.278	0.127	-0.051	0.153	-1.699	0.169	-0.023
	(2.45)**	(2.14)**	(1.10)	(0.29)	(0.64)	(2.05)**	(1.72)*	(0.16)
language	0.388	1.026	0.496	1.056	0.322	1.413	0.683	1.373
	(5.20)***	(10.02)***	(6.48)***	(8.34)***	(2.18)**	(3.23)***	(10.29)***	(13.22)***
Log(dist)	-0.408	-0.139	-0.448	-0.095	-0.613	0.057	-0.341	0.004
	(9.04)***	(2.48)**	(10.69)***	(1.63)	(7.40)***	(0.31)	(9.58)***	(0.08)
Schengen	0.168	0.065	0.277	0.599	-0.081	1.154	0.598	0.493
	(1.19)	(0.33)	(2.02)**	(2.56)**	(0.28)	(1.31)	(5.23)***	(2.71)***
Immig. pol		0.035		0.035		0.015		-0.338
		(7.85)***		(6.71)***		(0.87)		(3.33)***
Social exp		-0.290		0.175		2.411		0.236
		(2.25)**		(1.28)		(3.22)***		(6.98)***
Pop at dest		0.321		0.109		-0.131		0.033
		(9.66)***		(2.30)**		(0.83)		(7.51)***
Wages at dest		0.028		0.040		-0.020		0.069
		(3.70)***		(4.51)***		(0.75)		(9.32)***
Constant	3.750	-4.954	2.365	-6.119	1.388	-17.084	0.196	-6.701
	(6.92)***	(3.96)***	(4.02)***	(5.07)***	(1.20)	(2.99)***	(0.36)	(6.32)***

Table 2. Determinants of migration flows by skill level : OLS and Heckman regressions (2/2)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Total	Total	Total	Total	Low-skill	Low-skill	High-skill	High-skill
Observations	3608	3091	5760	4992	5760	4992	5760	4992
Dest dum	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Orig dum	yes	No	yes	No	yes	No	yes	yes
Method	OLS	OLS	Heckman	Heckman	Heckman	Heckman	Heckman	Heckman
Mills ratio	-	-	1.19	1.92	2.09	1.01	1.43	1.11
			(9.35)***	(7.65)***	(6.70)***	(10.6)***	(8.90)***	(8.75)***
R-squared	0.89	0.76						

Absolute values of robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Extracting the explained partial sum of squares using the results in column (1), we find that diaspora effects explain more than 71% of the observed variability in migration flows and over 80% of the explained variability of the model. This is a rather high level given that the fit of the regression is quite high, with R^2 amounting to 89%. Columns (5) and (6) report the results for the low-skill migrants while columns (7) and (8) report the results for the high-skill ones. The diaspora effect is higher for low-skill migrants as predicted in our model. This is due to the fact a large diaspora lowers the advantage higher levels of human capital generate in lowering migration and visa costs. The differential impact of diasporas on low-skill migration is again highly robust to alternative specifications (i.e. with and without destination country dummies) and to alternative estimation methods. A Wald test on the difference of coefficients of α_1 between low and high-skilled migrants (columns 5 and 7) shows that this difference is statistically significant at the 5% level. Note that the effects of distance and linguistic proximity are also higher for low-skilled than for the high-skilled migrants. The latter result reflects the fact that linguistic proximity increases the degree of transferability of skills and the ease of entry into the labor market for the low-skilled migrants.

Table 3 presents the instrumental variable estimates of equation (11) with three different sets of instruments. All sets pass the F-stat test for the strength of instruments and the Hansen J-test of no correlation with the error term at the 5% level. The results of the IV estimation lead to very similar coefficients for the impact of the diaspora on the migration flows. The decrease in significance is mainly caused by the increase in uncertainty due to the instrumentation procedure. Nevertheless, the quantitative and statistical significance of the diaspora remains. Therefore, we conclude that the strong effect of diasporas documented in OLS regressions is robust to the various econometric problems including selection bias and correlation of the diaspora with unobserved factors of the flows.

Table 3. Determinants of migration flows : IV estimation

	(1)	(2)	(3)
	Total	Total	Total
Lagged diasp	0.761	0.766	0.758
	(10.92)***	(11.09)***	(10.86)***
Col links	-0.051	-0.064	-0.045
	(0.26)	(0.32)	(0.23)
linguage	0.234	0.228	0.236
	(2.27)**	(2.22)**	(2.29)**
Log(dist)	-0.259	-0.253	-0.262
	(2.84)***	(2.78)***	(2.86)***
Schengen	0.160	0.161	0.160
	(1.11)	(1.11)	(1.11)
Constant	2.365	2.306	2.392
	(2.69)***	(2.64)***	(2.72)***
Orig Dum	Yes	Yes	Yes
Dest Dum	Yes	Yes	Yes
Method	IV	IV	IV
F-stat First stage	27.51	26.15	27.29
Hansen J-test (p-value)	0.128	0.0640	0.101
R2	0.883	0.882	0.883
Observations	3486	3486	3486

Absolute values of robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Instrument sets for M_{ij} in all columns include a dummy for bilateral guest worker agreement and a proxy for diaspora size in 1960. In column (1) , the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * Conf1_i$;In column (2) , the proxy is computed as $ln(pop_i \times immst_j/dist_{ij}) \times Conf2_i$;in column (3), the proxy is computed as $ln(pop_i \times immst_j/dist_{ij}) \times (conf1_i + Conf2_i)$.

5.2 Impact on Selection

The next question is on the determinants of the selection and the human capital (educational) composition of migrants and the specific role of diasporas in this process. Columns (1) to (4) in Table 4 report the results of the estimation of equations (12) and (13) for the skill ratio whereas columns (5) and (6) are estimated for the change in the skill ratio. Columns (1) and (2) are obtained using OLS whereas results in columns (3) to (6) are obtained with the Heckman two-step procedure without instruments.

Table 4. Impact of diaspora on selection (ratio high-skill/low-skill) level and change: OLS and Heckman

	(1)	(2)	(3)	(4)	(5)	(6)
	Skill ratio	Skill ratio	Skill ratio	Skill ratio	ΔSR	ΔSR
Lagged diasp	-0.171	-0.088	-0.194	-0.132	-0.143	-0.108
	(16.19)***	(8.47)***	(20.62)***	(11.83)***	(17.62)***	(11.47)***
Col. links	-0.042	-0.439	-0.022	-0.410	0.101	0.096
	(0.62)	(6.08)***	(0.32)	(5.21)***	(1.67)*	(1.46)
linguage	0.466	0.703	0.460	0.721	0.176	0.257
	(9.38)***	(11.03)***	(9.37)***	(11.68)***	(4.17)***	(4.95)***
Log(dist)	0.096	0.273	0.090	0.263	0.086	0.116
	(3.35)***	(10.17)***	(3.40)***	(9.96)***	(3.78)***	(5.25)***
Schengen	0.502	0.305	0.519	0.303	0.390	0.117
	(5.65)***	(3.14)***	(6.26)***	(2.97)***	(5.48)***	(1.37)
Immig pol		-0.014		-0.015		0.001
		(4.98)***		(5.52)***		(0.30)
Soc exp		-1.206		-1.253		-0.756
		(16.11)***		(20.12)***		(14.42)***
Pop. at dest		0.061		0.082		0.056
		(3.45)***		(4.58)***		(3.75)***
Wage at dest		0.044		0.045		0.035
		(9.86)***		(10.47)***		(9.78)***
Constant	-1.109	0.002	-0.734	0.257	-1.250	-0.563
	(1.16)	(0.00)	(1.32)	(0.34)	(2.54)**	(0.87)
Dest dum	Yes	No	Yes	No	Yes	No
Orig dum	Yes	Yes	Yes	Yes	Yes	Yes
Method	OLS	OLS	Heckman	Heckman	Heckman	Heckman
Mills			-0.380	-0.446	-0.10	-0.99
			(6.86)***	(7.37)***	(0.22)	(1.88)*
Obs	3604	3084	5760	4992	5760	4992
R-squared	0.60	0.45				

Robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Results in Table 4 show that the selection of migrants is also influenced by a large set of variables. Bilateral varaibles such as linguistic proximity, distance, the Schengen agreement and wage differentials favor the selection of high-skilled migrants. On the contrary, non-selective immigration policies and generous social expenditures lower the educational mix of the migrants which is in line with the results in Cohen and Razin (2008). More importantly, large diasporas exert a strong negative impact on the skill ratio of migration flows and attract low-skill migrants. Once again, this effect is robust to alternative specifications (presence or absence of destination dummies),

estimation methods (OLS and Heckman). More importantly, this strong result is maintained when we use the *change* in the skill ratio between 1990 and 2000 as the dependent variable instead of the *level* of the skill ratio.

From the results in column (1), we find that diaspora effects explain respectively 47% and 78% of the total and explained variability of the selection ratio in 2000. These numbers suggest that, compared to economic or other selection variables, diaspora effects are rather important. As stated earlier, the diaspora effect is complemetary to the generosity of family reunion programs. The size of the diaspora effect will be smaller in the absence of .reunification programs in the destination country and will be limited to lowering of migration costs through the network effects. These results imply that education-based selective migration policies are likely to have only moderate impact in countries hosting large diasporas unless the extent of family reunification programs are curtailed.

Table 5 reports the results of the IV estimation for the skill ratio of migration flows. Columns (1) through (3) look at the impact on the level of the ratio while Columns (4) through (6) investigate the impact on its change. Similar to the analysis of migration flows, we consider three different sets of instruments and the IV results confirm the negative impact of diasporas on the educational composition of migrant flows from the previous table.

5.3 Concentration

Our last question is on the determinants of the relative concentration of migrants of different skill levels and the role diasporas play. The structure of Table 6 is similar to the that of Table 4. Colonial links tend to favor a higher concentration of low-skill migrants compared to high-skill ones, while distance exerts the opposite effect. Once again, diaspora effects are found to be important for explaining the concentration levels and the effect is robust to alternative specifications, alternative estimation methods. And the results also hold for the change in the relative concentration index between 1990 and 2000.

Table 5. Impact of diaspora on selection (log high-skill/low-skill ratio): IV estimation

	(1) Log-skill ratio	(2) Log-skill ratio	(3) Log-skill ratio	(4) ΔLSR	(5) ΔLSR	(6) ΔLSR
Lagged diasp	-0.218	-0.207	-0.215	-0.215	-0.203	-0.212
	(3.01)***	(2.78)***	(2.95)***	(3.50)***	(3.22)***	(3.44)***
Col links	0.092	0.068	0.085	0.277	0.249	0.270
	(0.52)	(0.37)	(0.48)	(1.82)*	(1.61)	(1.77)*
linguage	0.469	0.459	0.466	0.238	0.226	0.235
	(5.41)***	(5.20)***	(5.37)***	(3.24)***	(3.04)***	(3.19)***
Log(dist)	0.057	0.067	0.060	0.016	0.028	0.019
	(0.75)	(0.87)	(0.79)	(0.25)	(0.43)	(0.30)
Schengen	0.536	0.536	0.536	0.414	0.415	0.414
	(6.38)***	(6.38)***	(6.38)***	(6.08)***	(6.09)***	(6.08)***
Constant	-0.468	-0.573	-0.501	-0.450	-0.567	-0.481
	(0.42)	(0.50)	(0.44)	(0.59)	(0.73)	(0.63)
Orig Dum	Yes	Yes	Yes	Yes	Yes	Yes
Dest Dum	Yes	Yes	Yes	Yes	Yes	Yes
Method	IV	IV	IV	IV	IV	IV
F-stat First stage	30.16	29.49	30.07	30.16	29.49	30.07
Hansen J-test (p-value)	0.974	0.244	0.812	0.574	0.562	0.747
R2	0.599	0.600	0.599	0.506	0.509	0.506
Observations	3486	3486	3486	3486	3486	3486

Absolute values of robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Instrument sets for M_{ij} in all columns include a dummy for bilateral guest-worker agreements and a proxy for diaspora size in 1960. In column (1) , the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * Conf1_i$.In column (2) , the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * Conf2_i$;in column (3), the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * (conf1_i + Conf2_i)$.

Table 6. Explaining relative concentration between high-skill and low-skill and change in relative concentration

	(1)	(2)	(3)	(4)	(5)	(6)
	Rel conc	Rel conc	Rel conc	Rel conc	ΔRC	ΔRC
Lagged diasp	-0.502	-0.294	-0.514	-0.347	-0.008	-0.008
	(5.87)***	(3.54)***	(9.67)***	(5.73)***	(16.05)***	(15.45)***
Col. links	-4.635	-7.085	-4.619	-7.008	-0.040	-0.043
	(4.68)***	(6.41)***	(10.69)***	(14.75)***	(9.93)***	(10.45)***
Language	0.338	0.373	0.321	0.369	-0.004	-0.005
	(0.84)	(0.78)	(1.09)	(1.02)	(1.58)	(1.75)*
Log(dist)	0.266	0.628	0.269	0.615	0.006	0.006
	(1.24)	(3.73)***	(1.69)*	(3.91)***	(3.78)***	(4.26)***
Schengen	-0.193	-0.076	-0.180	-0.068	0.002	0.001
	(0.50)	(0.16)	(0.36)	(0.11)	(0.49)	(0.26)
Pop. at dest		0.956		0.988		0.003
		(7.13)***		(9.33)***		(3.50)***
Immig pol		-0.014		-0.013		0.000
		(1.31)		(0.84)		(1.51)
Soc exp		-1.509		-1.573		0.002
		(4.38)***		(4.44)***		(0.52)
Wage at dest		0.217		0.217		0.001
		(7.69)***		(8.57)***		(4.68)***
Constant	5.607	-18.397	-3.240	-10.824	-0.037	-0.111
	(0.29)	(4.70)***	(1.19)	(2.77)***	(1.60)	(3.33)***
Dest dum	Yes	No	Yes	No	Yes	No
Orig dum	Yes	Yes	Yes	Yes	Yes	Yes
Method	OLS	OLS	Heckman	Heckman	Heckman	Heckman
Mills			-0.405	-0.680	-0.873	-1.684
			(1.07)	(1.94)**	(2.44)**	(6.12)***
Observations	3920	3367	5730	4966	5730	4966
R-squared	0.29	0.17				

Robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

5.4 Non-linear Effects

It is important to explore potential non-linear effects in our econometric specification, especially given the nature of the mechanism through which diasporas are expected to impact the characteristics of migration flows. Two sources of non-linearity can be expected. First, as the size of a diaspora expands, the marginal impact of an additional migrant could decline. Fortunately, the logarithmic specification accounts for this source of non-linearity.

Another potential issue is that diasporas below a certain size could be ineffective in lowering migration costs. In other words, smaller diasporas might lead to relatively high search costs for potential migrants, mitigating the positive effects reported in the previous tables. This possibility argues for presence of threshold effects in the impact of diasporas. In order to check the existence of such a threshold, we run rolling regressions of the following type:

$$ln(X_{ij}) = \alpha_0^s + \alpha_1^s \ln\left(\widetilde{M}_{ij}^s\right) + \alpha_2^s d_{i,j} + \gamma_j^s + \gamma_i^s + \epsilon_{ij}^s \tag{16}$$

with $X_{ij} = N_{ij}$ or S_{ij}, $\widetilde{M}_{ij}^s \equiv Max\left[M_{ij} - \underline{M}^s; 1\right]$ and $\underline{M}^s$, the threshold in diaspora size, varying between 0 to 7,500 migrants. Hence, we look at the impact of diasporas on the migration flows and on the migration selection, neutralizing the impact for diasporas whose size is lower than $\underline{M}^s$.

Given the distribution of the diaspora size, we roll over increments of 50 migrants up to 7,500 migrants. Note that by generating zero values for $\ln\left(\widetilde{M}_{ij}^s\right)$, one should expect the standard error of $\widehat{\alpha}_1^s$ to increase as $\underline{M}^s$ increases (the variability of $\ln\left(\widetilde{M}_{ij}^s\right)$ decreases in a nonlinear way). For instance, when $\underline{M}^s = 7500$, the proportion of zeros generated for $\ln\left(\widetilde{M}_{ij}^s\right)$ becomes higher than 89 percent. The estimation of $\widehat{\alpha}_1^s$ is then based on a very low number of observations[15]. This tends to inflate the standard errors of $\widehat{\alpha}_1^s$.

Figures 3 and 4 plot the evolution of the estimated $\widehat{\alpha}_1^s$ along with values of $\underline{M}^s$ with both estimations using Heckman two-step method. Both figures suggest that the impact of diasporas is slightly decreasing with the size of diasporas. The evolution over time of the estimated $\widehat{\alpha}_1^s$'s does not suggest the existence of a minimum threshold under which diasporas would be inefficient. Consequently, those results are fully consistent with the choice of a double log specification for models (11) and (13) and values of $\mid \widehat{\alpha}_1 \mid < 1$.

[15] Basically, the estimation will only rely on pairs of relatively large sending and receiving countries.

Figure 3. Estimating $\widehat{\alpha}_1^s$ with rolling regressions
Dependent = change in diaspora size

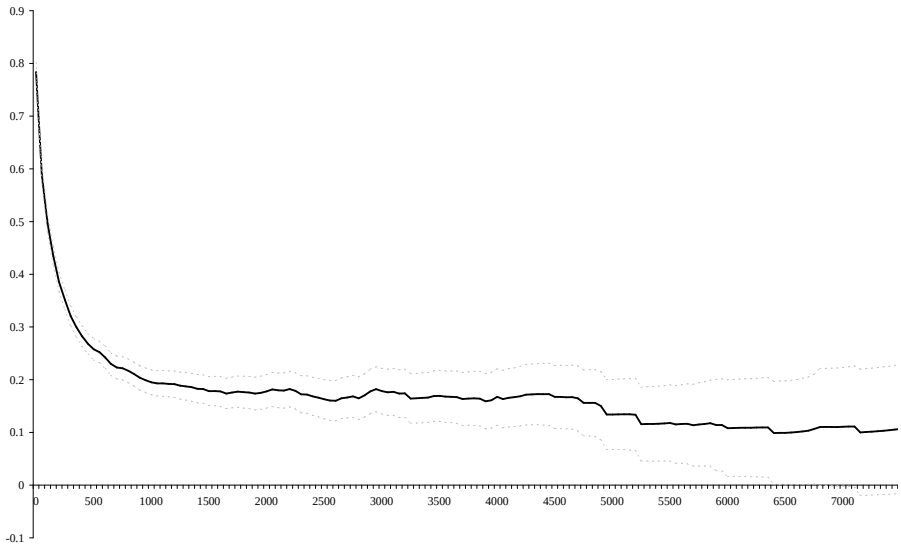
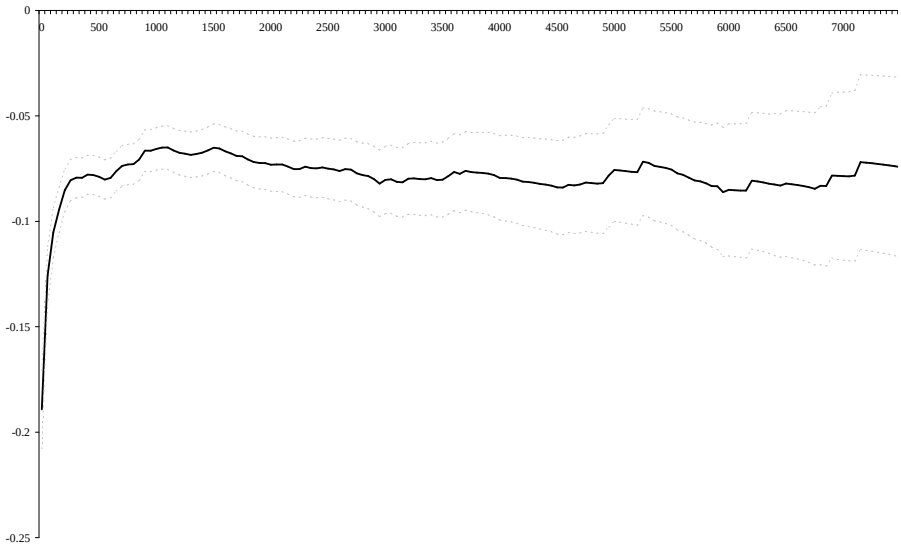


Figure 4. Estimating $\widehat{\alpha}_1^s$ with rolling regressions
Dependent = log skill-ratio



6 Conclusion

This paper explores the impact of existing diasporas on the number, skill composition and concentration of international migrants. We first develop a simple theoretical framework emphasizing the role of diasporas which operates through the lowering of both migration costs (due to information and assimilation difficulties) and visa costs (due to government policies). Diasporas lower migration costs through network effects and visa costs by increasing the probability of non economic migration through family reunification programs. These two effects increase the size of migration flows and reasonably reduce the selection of high-skill migrants. We then evaluate the implications of the theoretical predictions using new bilateral migration data by educational level. We estimate the role of existing networks in 1990 on the migration flows between 1990 and 2000, on their skill composition and on their relative concentration across educational levels. We account for potential problems related to the nature of the data and check the robustness of initial OLS estimates. First, we abstract from the bias induced by the log linearization of gravity models. Second, we account for the occurrence of a potential selection bias due to the large number of zeros in the country pairs. Third, we take into account a potential endogeneity problem of existing diasporas through instrumental variable estimations. Our results are extremely robust across estimation methods.

In short, we find evidence of a strong impact of existing diasporas. Regarding size, diasporas are by far the most important determinant of migration flows even after accounting for the usual variables affecting bilateral migration costs such as distance, colonial links and linguistic proximity. Extracting the explained partial sum of squares, we find that 71 percent of the observed variability of the migration flows is explained by diaspora effects. Regarding selection, diasporas are found to favor more the migration of the low-skill than migration of the highly skilled. It therefore exerts a strong negative impact on the selection of migrants. We find that diaspora effects explain 47% of the total variability of the selection ratio in 2000. Disregarding diaspora externalities but using much more detailed data on base wages and returns to skill (captured in our fixed effects), Grogger and Hanson (2008) find that, on average, wage differences explain 58 percent of the immigrant skill gap. This suggests that diaspora effects and wage differences leave little space for education-based selective policies in determining the quantity and quality of immigrants. Our results suggest that policies aiming at increasing the educational quality of the migrants might be highly constrained by the existing migrant's network. In the presence of large diasporas, more selective migration policies might fail unless family reunification programs are deeply reformed and limited. The same holds for policies that would aim to favor ethnic diversity of the migrants.

7 References

Bertolini, S. (2009), "Networks, Sorting and Self-selection of Ecuadorian Migrants", Paper presented at the second TOM Meeting, Louvain-La-Neuve, January.

Borjas, G (1987), "Self-selection and the earnings of migrants", American Economic Review, 77 (4), 531-53.

Borjas, G.J. (1994), "The economics of immigration", Journal of Economic Literature, 32, 1667-1717.

Borjas, G.J. (1995), "The economic benefits from immigration", Journal of Economic Perspectives, 9 (2), 3-22.

Borjas, G.J. (1999), Heaven's door: immigration policy and the American economy, Princeton University Press.

Carrington, W.J., E. Detragiache and T. Vishwanath (1996), "Migration with endogenous moving costs", American Economic Review, 86 (4), 909-30.

Chiquiar, D. and G.H. Hanson (2005), "International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States", Journal of Political Economy, 113 (2), 239-81.

Clair, G., G. Gaullier, Th. Mayer and S. Zignago (2004), "A note on CEPII's distances measures", Explanatory note, CEPII, Paris.

Cohen, A. and A. Razin (2008), "Skill composition of migration and the generosity of the welfare state: free vs. policy-restricted migration", Mimeo., Tel-Aviv University.

Docquier, F. and E. Lodigiani (2008), "International migration and business networks", Open Economies Review, forthcoming.

Docquier, F., O. Lohest and A. Marfouk (2007), "Brain drain in developing countries", World Bank Economic Review, 21, 193-218.

Docquier, F. and A. Marfouk (2006), "International migration by educational attainment (1990-2000)", in C. Ozden and M. Schiff (eds). International Migration, Remittances and Development, Palgrave Macmillan: New York (2006), chapter 5.

Docquier, F., B.L. Lowell and A. Marfouk (2007), "A gendered assessment of highly skilled emigration", Population and Development Review, forthcoming.

Friedberg, R.M. and J. Hunt (1995), "The impact of immigrants on the host country wages, employment and growth", Journal of Economic Perspectives, 9, 23-44.

Gao, T.(2003), "Ethnic Chinese Networks and International Investment: Evidence from Inward FDI in China", Journal of Asian Economics, 14, 611-629.

Gleditsch, P., M. Eriksson and M. Sollenberg (2002), "Armed Conflict 1946-2001: A New Dataset", Journal of Peace Research, 39 (5), 615-637.

Grogger, J and G.H. Hanson, 2008, "Income Maximisation and the selection and sorting of international Migrants, NBER Working Paper, No. 13821.

Harbom, L., E. Melander and P. Wallensteen (2007), "Dyadic Dimensions of Armed Conflict, 1946—2007", Journal of Peace Research, 45 (5), 697-710.

Helpman, E., M. Melitz and Y. Rubinstein (2007), "Estimating Trade Flows: Trading Partners and Trading Volumes", NBER Working Paper W12927.

Manski, C.F. (1993), "Identification of Endogeneous Social Effects: the Relection Problem", Review of Economic Studies, 60 (3), 531-42.

Massey, D.S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino and J. E. Taylor (1993), "Theories of international migration: Review and Appraisal," Population and Development Review, 19 (3), 431-466.

McFadden, D. (1984), "Econometric analysis of qualitative response models", in: Z. Griliches and M. Intriligator, eds., Handbook of Econometrics, Volume 2, Amsterdam. Elsevier/North-Holland.

McKenzie, D. and H. Rapoport (2007), "Self-selection patterns in Mexico-US migration: the role of migration networks", Review of Economics and Statistics, forthcoming.

Munshi, K. (2003), "Networks in the modern economy: Mexican migrants in the US labor market", Quarterly Journal of Economics, 118 (2), 549-99.

Pedersen, P.J., M. Pytlikova and N. Smith (2008),. "Selection and network effects– Migration flows into OECD countries 1990-2000", European Economic Review, 52 (7), 1160-1186.

Rapoport, H. and M. Kugler (2006), "Skilled Emigration, Business Networks and Foreign Direct Investment", CESifo Working Paper Series No. 1455.

Rauch, J. (2003), "Diasporas and development: Theory, Evidence and Programmatic Implications", Department of Economics, University of California at San Diego.

Rauch, J. and A. Casella (1998), "Anonymous Market and Group ties in International Trade", Journal of International Economics, vol 58(1):19-47.

Rauch, J. and V. Trindade (2002), "Ethnic Chinese Networks In International Trade", The Review of Economics and Statistics, MIT Press, vol. 84(1):116-130.

Razin, A. and E. Sadka (2004), " Welfare migration: Is the net fiscal burden a good measure of its economic impact on the welfare of the native-born population?", NBER Working Paper 10682.

Rosenzweig, M (2008), The global Migration of Skill, Paper presented at the Migration and Development Workshop, Lille, June.

Roy, A.D. (1951), "Some thoughts on the distribution of earnings", Oxford Economic Papers, 3 (2), 135-46.

Santos Silva, J.M.C. and S. Tenreyro (2006), "The Log of Gravity", Review of Economics and Statistics, 88 (4): 641-658.

Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press.

8 Appendix A - Robustness

Table A1 present the results obtained with diplomatic representation used as an additional instrument. Table A2 presents the results obtained with the Poisson regressions.

Table A1 : Determinants of migration flows
Heckman regressions with diplomatic representation as instrument

	(1)	(2)	(3)
	Total	Low-skill	High-skill
Lagged diasp	0.660	0.732	0.592
	(47.97)***	(25.65)***	(47.40)***
Col links	0.219	0.296	0.224
	(2.03)**	(1.42)	(2.37)**
linguage	0.477	0.315	0.658
	(6.71)***	(2.42)**	(10.25)***
Log(dist)	-0.501	-0.686	-0.387
	(12.04)***	(8.66)***	(10.71)***
Schengen	0.257	-0.090	0.610
	(2.00)**	(0.36)	(5.54)***
Constant	2.785	1.789	2.408
	(4.82)***	(1.44)	(4.19)***
Dest dum	Yes	Yes	Yes
Orig dum	Yes	Yes	Yes
Method	Heckman	Heckman	Heckman
Mills ratio	0.908	1.836	0.772
	(7.60)***	(6.77)***	(8.60)***
Diplomatic representation	0.202	0.171	0.010
	(2.36)**	(2.39)**	(1.08)
Observations	5610	5610	5610

Robust t statistics in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A2 : Determinants of migration flows (total and low skilled)
Poisson regressions

	(1)	(2)	(3)	(4)	(5)	(6)
	Total	Total	Low-skill	Low-skill	High-skill	High-skill
Lagged diasp	0.703	0.740	0.743	0.784	0.644	0.706
	(16.20)***	(22.06)***	(11.92)***	(15.09)***	(18.20)***	(22.70)***
Colonial links	-0.312	-0.375	0.183	0.169	-0.218	-0.305
	(1.65)*	(2.04)**	(0.67)	(0.58)	(1.39)	(2.35)**
language	0.298	0.369	-0.225	-0.266	0.522	0.551
	(2.53)**	(2.81)***	(1.45)	(1.48)	(4.86)***	(5.75)***
Log(distance)	-0.337	-0.186	-0.434	-0.341	-0.081	0.039
	(3.28)***	(2.37)**	(3.58)***	(3.85)***	(0.99)	(0.57)
Schengen	0.061	0.264	-0.628	-0.656	0.351	0.166
	(0.23)	(0.87)	(1.42)	(1.30)	(1.69)*	(0.73)
Immigr. policy		-0.053		0.090		0.021
		(0.30)		(0.39)		(2.83)***
Popul. at dest		0.284		0.271		0.316
		(5.39)***		(3.87)***		(6.25)***
Social exp		0.005		0.019		-0.022
		(0.52)		(1.82)*		(0.15)
Wages at dest		-0.023		-0.035		0.031
		(1.98)**		(2.48)**		(2.77)***
Constant	3.461	-2.251	3.219	-2.461	1.953	-6.049
	(3.06)***	(1.64)	(2.08)**	(1.47)	(2.35)**	(4.99)***
Dest dum	Yes	Yes	Yes	Yes	Yes	Yes
Orig dum	yes	No	Yes	No	Yes	No
Pseudo R2	0.955	0.945	0.963	0.960	0.848	0.875
Observations	5374	4649	4653	3974	5498	4762

Robust t-statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

9 Appendix B : Data sources

9.1 Migration data

- $M_{ij}(h)$: diaspora size by skill level h, defined as the number of people with skill level h ($h = 1, 2, 3$) aged 25+ born in country i and living in (OECD) country j.Source : Docquier, Lowell and Marfouk (2009).

- $N_{ij}(h)$: migration flows, skill level h, proxied by the change in $ln(M_{ij}(h))$ between 1990 and 2000.

9.2 Variables related to migration costs d_{ij}

- Geodesic distance in kms ($dist_{ij}$). Source : Clair, Gaullier, Mayer and Zignago (2004).

- Colonial Links : Dummy variable capturing whether there is a colonial link after 1945 between i and country j. Source : Clair, Gaullier, Mayer and Zignago (2004).

- Linguistic proximity : Dummy variable capturing a common language between i and country j. Source : Clair, Gaullier, Mayer and Zignago (2004).

- Schengen agreement : dummy variable taking 1 if both countries are subject to the the Schengen agreement between European countries. Source: European Commission.

- Skill price w_j : Estimates obtained from a log wage equation based on the US New Immigrant Survey (estimated across 3,994 workers aged 22+ when they last worked in their home country and who reported a wage at their last job). The specification included the worker's age and its square and the log of the year when the wage was reported, gender, and schooling in years. The predicted skill price by country in the data set is the hourly wage for a male worker with 12 years of schooling at age 40 for the year 2000. Source : Rosenzweig (2008)

- Social expenditure as a share of GDP, A_j^1. Source OECD

- Degree of selective immigration policy, A_j^2: captured by the share of refugees in the total number of migrants, year 1980 or 1990. Source : United Nations Population Division.

- Population size in country j. Source : United Nations Population Division.

9.3 Diplomatic representation (for Heckman estimation)

- Diplomatic representation : dummy variable capturing type of diplomatic representation of country j in country i prevailing in 1990. This variable can take 4 different values capturing the strength of the diplomatic representation. Source: Correlates of War Diplomatic Exchanges, version 2006.1.

9.4 Instruments of Diasporas (IV estimation)

- Guest Worker agreement: dummy variable taking 1 if there was a bilateral guest worker agreement in the 50's and 60's between country i and country j facilitating the migration of workers from country i. Own computations.

- Proxies for potential diaspora M_{ij} prevailing in 1960:

$$ln(pop60_i * immst_j/dist_{ij}) * Conf_i$$

where $pop60_i$ is the population size in 1960 (source: United Nations Population Division), $immst_j$ is the stock of migrants in country j in 1960 (source: United Nations Population Division) and $Conf_i$ is a variable capturing the number of conflicts in country i between 1946 and 1960. $Conf_i$ can be measured in three ways. First, $Conf_i = conf1_i$ where $conf1_i$ is the number of armed conflict with death numbers comprised between 25 and 999. Second, $Conf_i = conf2_i$ where $conf2_i$ is the number of armed conflict with death numbers over 999. Third, $Conf_i = conf1_i + conf2_i$. Source : We use the PRIO armed conflicts database (version 4-2008), a conflict-year data set with information on armed conflicts where at least one party is the government of a state in the time period 1946-2007. A description of this data set can be found in Gleditsch et al. (2002). Changes introduced in the updated version 4-2008 are described in Harbom et al. (2007).